Chapter 19

Emerging Patterns as Structural Alerts for Computational Toxicology

Bertrand Cuissart, Guillaume Poezevara, Bruno Crémilleux

Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen, University of Caen Basse-Normandie, CNRS UMR 6072

Alban Lepailleur, Ronan Bureau

Centre d'Etudes et de Recherche sur le Médicament de Normandie, University of Caen Basse-Normandie, UPRES EA 4258 - FR CNRS 3038

To appear in the book entitled "Contrast Data Mining: Concepts, Algorithms and Applications", edited by Guozhu Dong and James Bayley, publisher Taylor and Francis Group, 2012

19.1 Introduction

Thanks to significant advances on both the algorithmic and the practical sides, mining graph data has turned into a key domain of data mining. Various domains use graphs to model their data and graph patterns have widely demonstrated their potential, especially in the field of chemoinformatics where chemical structures are commonly modeled as graphs. Computational toxicology, which aims at studying toxicity by using computer tools, is a typical example of an important field for developing graph mining methods. Even though there already exist useful tools such as Derek [12] that rely on fragments for assessing the toxic behavior of molecules, these methods suffer from two limitations [5]: (i) there is a lack of objectivity when a human expert assesses the level of toxicity caused by a molecular fragment and (ii) there is no decision rule based on the conjunction of two or more molecular fragments. Thus, there is a strong need of methods that can extract conjunctions of molecular fragments whose occurrences demonstrate relationships with a toxic behavior.

The chapter meets this need by designing a method, based on the no-

tion of *emerging pattern* [3], called the *Frequent Emerging Molecular Pattern (FEMP)* [10, 11]. Given a chemical dataset partitioned into two classes (e.g. toxic molecules and non-toxic ones), a FEMP is a conjunction of molecular fragments such that: (i) its frequencies between the classes are sufficiently different and (ii) its frequency in the target class is high enough to be significant to support a further use. In chemoinformatics, this notion positively answers the need of an automatic and understandable method for extracting the conjunctions of fragments related to a given behavior: FEMPs have already demonstrated their usefulness in computational ecotoxicology [7].

Section 1.2 explains our methodological contributions, i.e. (i) the FEMPs, their computation, and (ii) a condensed representation summarizing the extracted information. Section 1.3 gives the key results of an experimental study, where we quantitatively assess the effectiveness of the FEMPs as structural alerts in computational toxicology. Section 1.4 provides a thorough chemical analysis of the information brought by the extracted FEMPs. That chemical analysis represents a solid qualitative reasoning that advocates and demonstrates the advantages of the use of the FEMPs in computational toxicology.

19.2 Frequent Emerging Molecular Patterns as Potential Structural Alerts

This section introduces the notion of a *Frequent Emerging Molecular Pattern (FEMP)*. Here, we stress on the intuitions and the key ideas, and illustrate them by the example in Figure 1.1. Formal definitions and proofs of the results are given in [11]. Sections 1.3 and 1.4 will show that FEMPs are at the core of the chemical information discovered by data mining processes.

19.2.1 Definition of Frequent Emerging Molecular Pattern

The left side of Figure 1.1 displays molecular structures in the usual manner: *2D molecular graphs*. *Graphs* are frequently used to model elements having relationships – the edges of the graphs represent relationships. Then, a molecular structure is depicted as a set of elements, the atoms, that interact by means of edges, the chemical bonds. An element of a molecular graph is *labeled* with the atomic number it represents, while the label of an edge indicates the type of the chemical bond. In Figure 1.1, the chemical dataset is partitioned in two parts: toxic molecules and non-toxic ones.

The right side of Figure 1.1 displays *molecular fragments*; a molecular fragment represents a part of a molecule. A fragment is said to *occur* within a molecule if there is an embedding of the fragment in the molecule that satisfies both the relational structure of the fragment (the presence and the

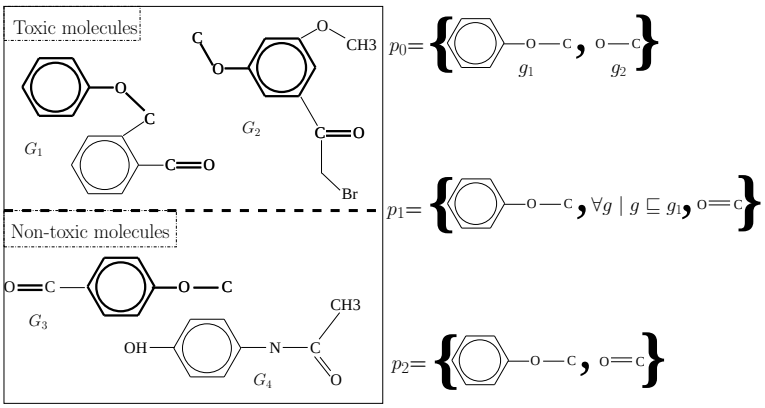


FIGURE 19.1: An illustrative chemical dataset

absence of every edge) and the labeling scheme of the edges. As an example, the embedding of the fragment g_1 is shown in bold on the molecular graphs. The *frequency of a fragment* in a chemical dataset quantifies the portion of molecules of the dataset where the fragment occurs. For instance, g_1 occurs in 100% of the toxic molecules and 50% in the non-toxic ones (cf. Figure 1.1). A *molecular pattern* is defined to be a set of molecular fragments. The *length* of a molecular pattern corresponds to the number of fragments it contains. A molecular pattern *occurs* in a molecule if each one of its fragments occurs in the molecule. The *frequency of a pattern* in a chemical dataset quantifies the portion of molecules of the dataset where the pattern occurs. In Figure 1.1, the pattern p_0 has a frequency of 50% in the non-toxic molecules.

In order to automatically discover structural alerts, it appears to be highly appropriate to look for contrasts between toxic and non toxic molecules. When a molecular pattern sufficiently occurs within the toxic molecules and has a frequency which significantly increases from the non-toxic molecules to the toxic ones, then it stands as a potential structural alert related to the toxicity.

The notion of a Frequent Emerging Molecular Pattern embodies this natural idea by using the growth-rate measure. When a dataset is partitioned between targeted examples and non-targeted ones (also called "classes"), the *growth-rate of a pattern* p is defined to be the ratio between the frequency of p in the target class over its frequency outside the target class. Following our example, the growth-rate of a molecular pattern is obtained by dividing its frequency in the toxic molecules by its frequency in the non-toxic ones. As usual when the denominator is equal to zero (and the numerator is different of zero), the value of the growth-rate is denoted with the infinity symbol, ∞. An *emerging pattern* [3] is a pattern whose growth-rate value exceeds a threshold given by the user. For example, in Figure 1.1, the growth-rate value of the molecular pattern p_0 is equal to 2. Thus, as soon as the minimum threshold

is set to less than 2, p_0 is an emerging pattern, from the non-toxic molecules to the toxic molecules. We can now state the definition of a FEMP:

Definition 1 (Frequent Emerging Molecular Pattern (FEMP)) *Let $\mathcal{G}$ be a chemical dataset whose molecules are partitioned into two classes. Given a frequency threshold f_{min} and a growth-rate threshold ρ_{min}, a molecular pattern p is a* frequent emerging molecular pattern *if its frequency in the target class is greater or equal than f_{min} and its growth-rate from the non-target class molecules to the target class is greater or equal than ρ_{min}.*

To simplify, from now on the property "being a FEMP in the chemical dataset $\mathcal{G}$ partitioned according to a classification and given a frequency threshold f_{min} and a growth-rate threshold ρ_{min}" will be abbreviated as "being a FEMP".

19.2.2 Using RPMPs as Condensed Representation of FEMPs

In practice, FEMPs are often numerous and include redundant information. This section deals with this issue by proposing a condensed representation of the FEMPs – the Representative Pruned Molecular Patterns (RPMP).

The growth-rate of a molecular pattern is computed from its frequencies in the classes in the chemical dataset. It implies that the property of being a FEMP only relies on its *extent*, the set of molecules of the input dataset in which the molecular pattern occurs. Patterns can be condensed by using specific forms of patterns such as the closed patterns [9, 13]; by condensation we mean that all FEMPs can be regenerated (from the condensed patterns) with their exact values of growth-rate and frequency. A *closed pattern* is a pattern from which no element can be added without decreasing its extent. Given any extent of the input dataset, there is at most one closed graph pattern corresponding to this extent. Relying on this property, one can condense a set of FEMPs by retaining only the related *closed* FEMPs. This significantly reduces the number of patterns without loosing information.

A molecular fragment may occur in another fragment, in the same way as a molecular fragment occurs in a molecule. As a consequence, the relationship "occurs in" induces a partial order between the molecular structures, depicting either a molecule or a fragment. The fact that g_1 occurs in g_2 is denoted $g_1 \sqsubseteq g_2$. As it is a partial order, the relation $\sqsubseteq$ is transitive and it follows that the extent of a fragment is included in the extent of any of its "subfragments". For example, the benzene ring () is a subfragment of the fragment g_1 of Figure 1.1, and its extent in the chemical dataset $\mathcal{G}$, namely $\{G_1, G_2, G_3, G_4\}$, contains the extent of g_1, namely $\{G_1, G_2, G_3\}$. Thus, one can add a new fragment to a molecular pattern without decreasing its extent, as long as the added fragment is a subfragment of an element of the pattern. For example, adding the benzene ring to the molecular pattern p_o in Figure 1.1 will not change the extent of p_0 because the benzene is a subfragment of g_1, which is

an element of p_0. As a consequence, a closed molecular pattern always contains all the subfragments of any of its fragments.

In practice, closed patterns tend to be long patterns and a large portion of their fragments are subfragments of another bigger fragment. These subfragments have no meaning and can be removed without loss of information. By pruning any fragment of a closed pattern p which is a subfragment of another fragment of p, we get a shorter representation of the closed pattern. We call the resulting patterns *Representative Pruned Molecular Patterns (RPMPs)*. There is a one-to-one correspondence between the RPMPs and the closed molecular patterns, and this correspondence preserves the extent [11]. For example, in Figure 1.1, the pattern p_1 is the closed pattern sharing the same extent as the pattern p_0. Not only p_1 contains g_1, but also all its subfragments. p_2 is the RPMP associated to the closed pattern p_1; p_2 has been built by removing any fragment of p_1 that is a subfragment of another fragment of p_1. p_2 enables us to provide a meaningful and understandable representation of p_1.

Definition 2 (Representative Pruned Molecular Pattern (RPMP)) *Let $\mathcal{G}$ be a chemical dataset. A molecular pattern p is a* representative pruned molecular pattern *if the molecular pattern obtained by adding all the subfragments of the elements of p is a closed molecular pattern in $\mathcal{G}$.*

Given a chemical dataset partitioned between targeted and non-targeted molecules, a minimum frequency threshold and a minimum growth-rate threshold, we have designed a method to mine the set of the Representative Pruned Molecular Patterns that are Frequent Emerging Molecular Patterns. This is discussed next.

19.2.3 Notes on the Computation

This section gives a sketch of our method for computing RPMPs. Let $\mathcal{G}$ be a given a dataset partitioned into targeted molecules and non-targeted ones. Recall that a FEMP is a molecular pattern whose frequency amongst the targeted molecules exceeds a given frequency threshold f_{min}. Since the relation "occurs in" is transitive, The frequencies, in the target class of molecules, of all molecular fragments in a FEMP must also exceed f_{min}. Our method uses this property by firstly extracting these frequent fragments. Second, it describes each molecule of the dataset by indicating for each of the frequent fragments whether it occurs or not in the molecule. Third, relying on this new description, one extracts the FEMPs by using an existing method dedicated to discover the emerging patterns.

The FEMPs and their associated RPMPs are computed by integrating three existing tools: Gaston [8], MicMac [13] and VFLib [2]. First, frequent fragments amongst the targeted molecules are extracted using the graph mining tool Gaston. Gaston computes the extent of a frequent fragment amongst the targeted molecules. We have updated it in order to simultaneously provide the extent of a frequent fragment in the whole input $\mathcal{G}$. The

result of the step is a dataset $\mathcal{D}$ which is a description of $\mathcal{G}$ based on the occurrences of the frequent fragments amongst the targeted molecules. Then, MicMac mines from $\mathcal{D}$ the closed patterns that also are frequent emerging patterns. Finally, the RPMPs are obtained by pruning these closed patterns: a fragment is removed from a closed pattern p as soon as it is a subfragment of another fragment of p. This step requires to perform subgraph isomorphism tests; they are done using an implementation based on the functionalities provided by the graph matching library VFLib.

19.2.4 Related Work

Several methods have been designed for discovering graphs that are correlated to a given class. All these algorithms operate on a graph dataset partitioned into two classes, targeted examples and non-targeted ones.

Molfea [6] relies on a level-wise algorithm. It extracts the *linear subgraphs(chains)* which are frequent amongst the targeted examples and infrequent amongst the non-targeted ones. However, the restriction to linear subgraphs disables a direct extraction of the fragments containing a branching point or a ring, as the benzene.

Moss [1] is a program dedicated to discover frequent fragments by mining molecular graphs; it can be extended to find the *discriminative fragments*. Given two frequency thresholds f_M and f_m, a discriminative fragment corresponds to a connected fragment whose frequency is above f_M amongst the targeted molecules and below f_m amongst the non-targeted ones. This definition differs from the usual notion of emergence which is based on the growth-rate measure as introduced in the previous section. Note that the set of the discriminative fragments according to the thresholds f_M and f_m does not contain the whole set of the FEMPs having a growth rate higher than f_M/f_m or any other given growth rate threshold. Moreover, such fragments only correspond to patterns of length 1.

Another work has been dedicated to the discovery of the *contrast fragments* [14]. A contrast fragment is a fragment that occurs in the targeted examples and never occurs in the non-targeted ones. Although this notion is very interesting, it requires a lot of computation. To the best of our knowledge, the calculus is limited to graph datasets containing one targeted example and the mining of a molecule exceeding 20 atoms brings up a significant challenge.

19.3 Experiments in Predictive Toxicology

This section aims to experimentally assess the utility of the Frequent Emerging Molecular Patterns (FEMPs) for predictive toxicology. Following the results shown in the previous section, we use Representative Pruned Molec-

ular Patterns (RPMPs) as a condensed representation of the FEMPs. For the sake of simplicity, in the following by a RPMP we mean a RPMP which is also a FEMP. In this section, first, the chemical dataset and the experimental setup are detailed, then the potential of the FEMPs is assessed by examining whether they retain their properties outside a learning set. Finally, quantitative results about the RPMPs in predictive toxicology are provided.

19.3.1 Materials and Experimental Setup

Chemical Dataset

Data were obtained from the *EPA Fathead Minnow Acute Toxicity Database* (EPAFHM) [4]. The data were collected by the Environment Protection Agency of the United States. EPAFHM has already been used for expert systems in computational toxicology [15]. The chemical dataset used here include molecules selected from EPAFHM based on the LC50 value associated to the molecules[1] – we selected the molecules known as very toxic or non-toxic. The dataset includs 297 molecules, partitioned according to their level of toxicity: 74 molecules are very toxic and 172 are non-toxic.

Experimental Setup

Results given in this section are obtained from averages over a five-folds cross-validation scheme: The dataset was randomly shuffled and then divided into five folds, such that each fold preserves the initial ratio between very toxic molecules and non-toxic ones. Each fold is successively the *test set*, with the union of the four other folds forming the *learning set*. A learning set averages 196.8 molecules (59.2 very toxic and 137.6 non-toxic) whereas a test set averages 49.2 molecules (14.8 very toxic and 34.4 non-toxic).

By definition, the property of "being a RPMP" relies on two thresholds: a minimum frequency value and a minimum growth-rate value. Throughout this experiment, the minimum frequency threshold is set to 8% (i.e., a pattern has to appear in 5 very toxic molecules to be extracted) and the minimum growth-rate threshold varies. With this minimum frequency threshold, 104.2 frequent fragments are extracted on average from a learning set (consisting of the very toxic molecules), and these frequent fragments contain on average 5.7 atoms. The number of RPMPs decreases from 318 (when the growth-rate value is 2) to 43.3 (when the growth-rate value is set to ∞). The RPMPs are mostly conjunctions of several molecular fragments and their average length is between 2 and 3 fragments, whatever the growth-rate threshold value is.

[1] The Lethal Concentration 50 (LC50) of a molecule indicates the concentration that kills half of a population of fishes; for the sake of simplicity, the term "toxicity" is used even if the LC50 indicates the ecotoxicity of a molecule.

Growth-rate threshold	Category (%) i	ii	iii	iv	v
2	25.2	13.1	30.0	16.7	14.6
5	24.4	12.0	32.1	11.1	20.1
10	27.6	8.7	34.1	6.6	22.8
25	33.7	1.2	31.3	3.0	30.5
∞	34.8	0.0	29.3	2.6	33.1

TABLE 19.1: The RPMPs outside the learning set.

19.3.2 Generalization of the RPMPs

Generalization of the Properties of the RPMPs

As previously seen, "being a RPMP" relies on two key properties: (i) a RPMP is frequent enough to be representative and to ensure further uses and (ii) its growth-rate value conveys a relation between the RPMP and the toxicity. This section assesses whether these key properties can be generalized outside the learning set. For that purpose, we follow the cross-validation scheme and we examine the behavior of every RPMP in the test set related to the learning set it has been extracted from. By examining its extent in the test set, each RPMP is classified into one of the following five exclusive categories: (i) the RPMP meets both the frequency threshold and the growth-rate threshold, (ii) it only meets the frequency threshold, (iii) it only meets the growth-rate threshold, (iv) it meets neither the frequency threshold nor the growth-rate one, (v) it does not occur in the test set.

Table 1.1 gives the portions of the RPMPs in each category for several growth-rate thresholds. The sum of the portions of the first three categories shows that two-thirds of the RPMPs still meet the frequency threshold or the growth-rate one in a test set. By comparing results in categories (ii) and (iii), one note that a RPMP more often meets the growth-rate threshold than the frequency one (a half of the RPMP meets the growth-rate threshold whereas only a third still meets the frequency one). Taken together, these results indicate that the key properties associated to a RPMP are satisfied outside the learning set.

The RPMPs for Predicting Toxicity of Molecules

In order to quantitatively assess the RPMPs in predictive toxicology, the following decision rule has been implemented: *a molecule is classified as very toxic if it contains at least one RPMP*. Table 1.2 displays the results by using such a classification rule on the related test set. The first column gives the value of the growth-rate that has been used for extracting the RPMPs. The *coverage rate* indicates the portion of the molecules of a test set that contains at least one RPMP. The *coverage contrast* corresponds to the ratio of the coverage rate amongst the very toxic molecules over the coverage rate amongst the non-toxic ones. *TP* (i.e., True Positive) displays the portion of very toxic

Growth-rate threshold	Length (l)	Coverage		Succes (%)		
		rate (%)	contrast	TP	TN	OV
2	**l≥1**	71.1	1.54	94.5	38.9	55.69
	l=1	64.2	1.84	94.5	48.8	62.6
5	**l≥1**	44.7	3.0	83.7	72.0	75.6
	l=1	37.8	4.03	79.7	80.2	80.08
10	**l≥1**	34.1	4.9	77.0	84.3	82.11
	l=1	26.4	6.06	63.5	89.5	81.7
25	**l≥1**	23.1	6.5	56.7	91.2	80.89
	l=1	13.0	6.97	32.4	95.3	76.42
∞	**l≥1**	20.3	5.97	48.6	91.8	78.86
	l=1	9.3	8.36	24.3	97.0	75.2

TABLE 19.2: Prediction of the toxicity of a molecule thanks to the RPMPs.

molecules that are correctly processed by the decision rule and *TN* (i.e. True Negative) is the ratio of non-toxic molecules correctly processed. *OV* indicates the overall success rate of the decision rule.

Results show that such a decision rule is able to reach fair overall success rates, greater than 80%. Moreover the contrast values indicate that the decision rule is more often triggered in the very toxic molecules than in the non-toxic ones; such a result indicates the reliability of the process. Table 1.2 also provides the results obtained by using only the RPMPs of length 1. We see that the portions of very toxic molecules that are correctly classified (i.e. *TP*) are significantly higher by using the whole set of RPMPs instead of RPMPs of length 1. Thus, one concludes that there exist conjunctions of non-emerging molecular fragments that have an influence on the toxic behavior of a molecule. Relationships between the chemical composition of some RPMPs and their effect on toxicity are discussed in the following section.

19.4 A Chemical Analysis of RPMPs

RPMPs have the advantage to support a chemical analysis. This section describes such an analysis that gives valuable new information for structure-toxicity relationships. The analysis is carried out according to two chemical functions or groups, the alkyl chains and the aromatic groups.

Alkyl Chains

A first illustration deals with the impact of the order associated to the alkyl chains (the fragments are ordered by their number of atoms). The corresponding discovered patterns (cf. Table 1.3) show a clear relation between the

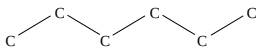
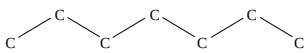
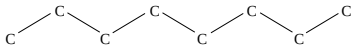
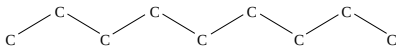
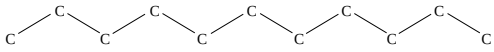
Growth-rate	Molecular fragment
2.7	
6.9	
11.5	
13.8	
∞	

TABLE 19.3: Growth-rate values of the alkyl chains according to their order

growth-rate values and their orders of the fragments. The meaningful order of the alkyl chains begins for C6 (6 carbons, growth-rate of 2.7), it increases strongly for C7 (growth-rate of 6.9) to reach a maximum value for C11 (a growth-rate of ∞). It is well known that the hydrophobicity of an alkyl chain correlates with its order and that hydrophobicity of a fragment favors a toxic behavior. The above analysis shows that the growth-rate values match the chemical knowledge on toxicity.

Aromatic Groups

The second illustration is related to the aromatic groups. These groups have a strong impact on the toxicity of chemicals. Our analysis shows that the nature of the substituents on the aromatic ring plays a major role on toxicity.

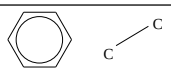
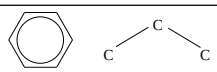
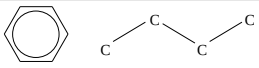
Growth-rate	Molecular fragments
3.01	
3.06	
10.7	
20.7	

TABLE 19.4: Association between alkyl chains and aromatic groups

The first example is the combination between an aromatic group and an

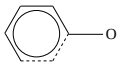
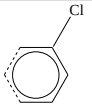
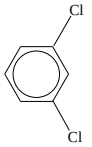
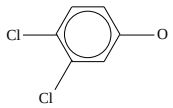
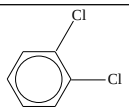
Chlorines in the benzene function		Chlorines in the phenol function	
Growth-rate	**Molecular fragments**	**Growth-rate**	**Molecular fragments**
2.08		3.64	
7.66		13.8	
16.1		∞	
∞			

TABLE 19.5: Addition of chlorines

alkyl chain. An aromatic group alone has a growth-rate value of 3. Associated with a C2 alkyl chain, we do not observe a modification of the growth-rate value but the growth-rate increases strongly for C3 and C4 alkyl chains (cf. Table 1.4).

The second example corresponds to the chlorinated benzenes or chlorinated phenols (cf. Table 1.5, the dotted lines depict a flexibility for the nature of the last atom associated to the aromatic feature). Chlorine atoms on aromatic groups lead to increase the toxicity. This point clearly appears in our study. The addition of one chlorine increases the growth-rate by a factor around 4. The addition of two chlorines increases the growth-rate by a factor 8, reaching a maximum value with two chlorines in ortho positions on the aromatic group. We observe the same evolution for the phenol functions.

The third example concerns the combination between an alkene function and an aromatic group (cf. Table 1.6-A). The growth-rate is maximum (i.e. ∞), showing the potential high toxicity of the association. One can note that the alkene function alone has a high growth-rate value (10.73).

The last example deals with the association between a carbonyl function and an aromatic group (cf. Table 1.6-B). In this case, we observe no significant evolution of the growth-rate values compared to the patterns without this function (see the alkyl chains and the aromatic group in Table 1.4). So, the impact of the carbonyl function on the toxicity is not characterized.

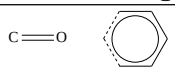
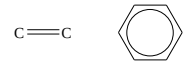
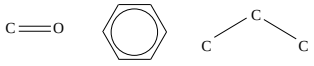
A- Alkene and aromatic group		B- Carbonyl function and aromatic group	
Growth rate	**Molecular fragments**	**Growth rate**	**Molecular fragments**
10.73		3.06	
∞		10.35	
		16.01	

TABLE 19.6: Combination with an aromatic group

19.5 Conclusion

In this chapter, we have defined the notion of a frequent emerging molecular pattern, such a pattern being useful in chemoinformatics. We have shown that the whole set of frequent emerging molecular patterns can be condensed by means of their related representative pruned molecular patterns. An experimental study has been carried out on a chemical dataset. This study has indicated the effectiveness of using the information provided by the occurrences of frequent emerging molecular patterns in predictive toxicology: a decision rule based on such patterns can distinguish between a very toxic molecule and non-toxic one in 80% of the cases. Besides, it has been shown in Section 1.4 that the evolution of the growth-rate values associated to each Representative Pruned Molecular Patterns gives new keys to understand the impact of an atom, a group of atoms or a chemical function on the toxicity of a chemical derivative. These results strongly advocate the potential of this new approach for computational toxicology.

Bibliography

[1] Christian Borgelt and Michael R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *IEEE International Conference on Data Mining (ICDM)*, pages 51–58, 2002.

[2] Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26:1367–1372, 2004.

[3] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 43–52, 1999.

[4] EPAFHM. *Environement Protection Agency Fathead Minnow Acute Toxicty*, version of 2008. http://www.epa.gov/med/Prods_Pubs/fathead_minnow.htm.

[5] Luis G. and Valerio Jr. In silico toxicology for the pharmaceutical sciences. *Toxicology and Applied Pharmacology*, 241(3):356 – 370, 2009.

[6] Stefan Kramer, Luc De Raedt, and Christoph Helma. Molecular feature mining in hiv data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 136–143, New York, NY, USA, 2001. ACM.

[7] Sylvain Lozano, Guillaume Poezevara, Marie-Pierre Halm-Lemeille, Elodie Lescot-Fontaine, Alban Lepailleur, Ryan Bissell-Siders, Bruno Cremilleux, Sylvain Rault, Bertrand Cuissart, and Ronan Bureau. Introduction of jumping fragments in combination with qsars for the assessment of classification in ecotoxicology. *Journal of Chemical Information and Modeling*, 50(8):1330–1339, 2010.

[8] Siegfried Nijssen and Joost N. Kok. A quickstart in frequent structure mining can make a difference. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 647–652, 2004.

[9] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *International Conference on Database Theory*, pages 398–416, 1999.

[10] Guillaume Poezevara, Bertrand Cuissart, and Bruno Crémilleux. Discovering emerging graph patterns from chemicals. In *International Symposium on Foundations of Intelligent Systems (ISMIS)*, pages 45–55, 2009.

[11] Guillaume Poezevara, Bertrand Cuissart, and Bruno Crémilleux. Extracting and summarizing the frequent emerging graph patterns from a dataset of graphs. *J. Intell. Inf. Syst.*, 37(3):333–353, 2011.

[12] JE Ridings, MD Barratt, R Cary, CG Earnshaw, CE Eggington, MK Ellis, PN Judson, JJ Langowski, CA Marchant, MP Payne, WP Watson, and TD Yih. Computer prediction of possible toxic action from chemical structure: An update on the DEREK system. *Toxicology*, 106(1-3):267–279, JAN 8 1996.

[13] Arnaud Soulet and Bruno Crémilleux. Adequate condensed representations of patterns. *Data Min. Knowl. Discov.*, 17(1):94–110, 2008.

[14] Roger Ming Hieng Ting and James Bailey. Mining minimal contrast subgraph patterns. In *SIAM International Conference on Data Mining*, pages 638–642. SIAM, 2006.

[15] GD. Veith, B. Greenwood, RS. Hunter, GJ. Niemi, and RR. Regal. On the intrinsic dimensionality of chemical structure space. *Chemosphere*, 17(8):1617–1644, 1988.